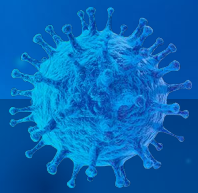
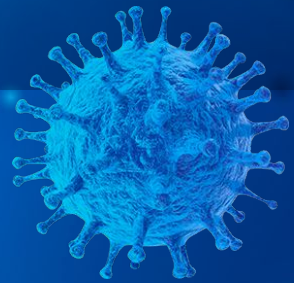
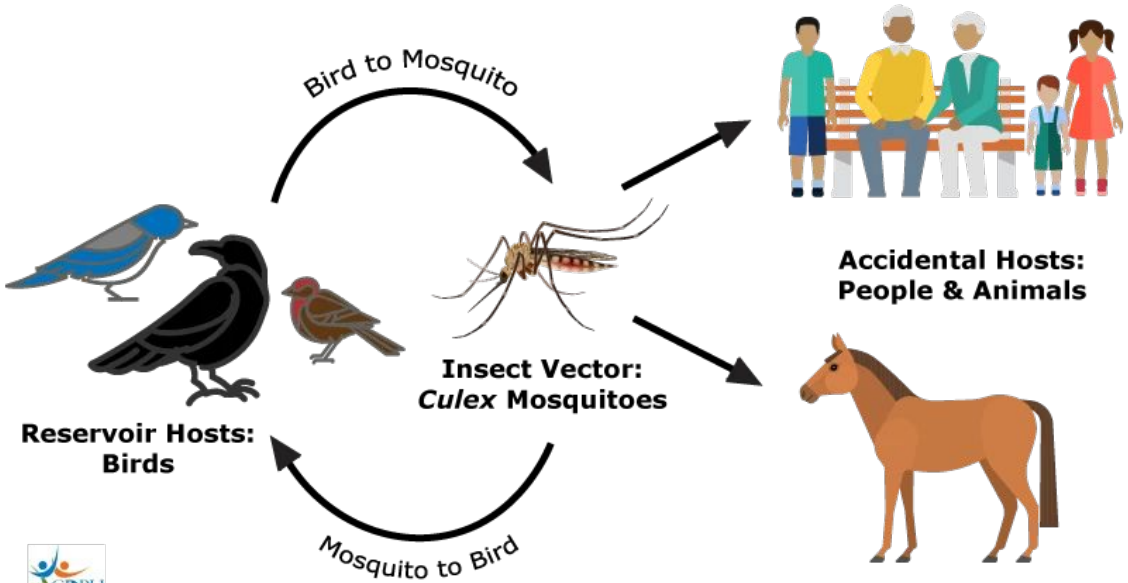
West Nile Virus

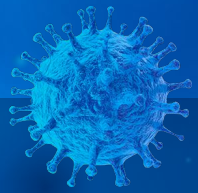


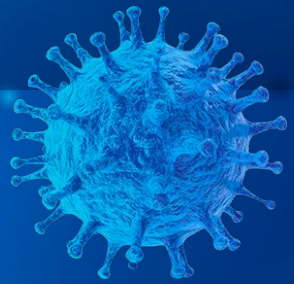


West Nile Virus Transmission Cycle



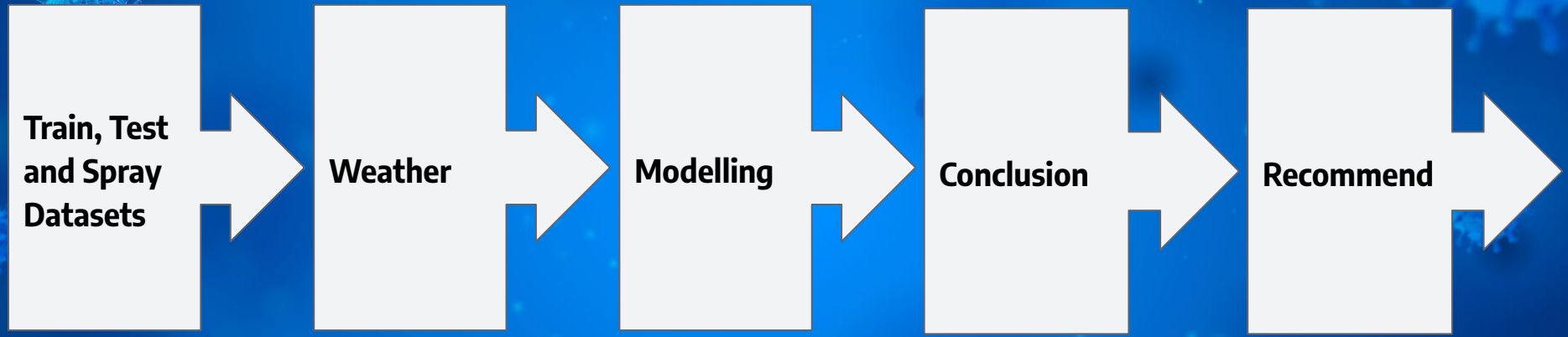
Problem Statement



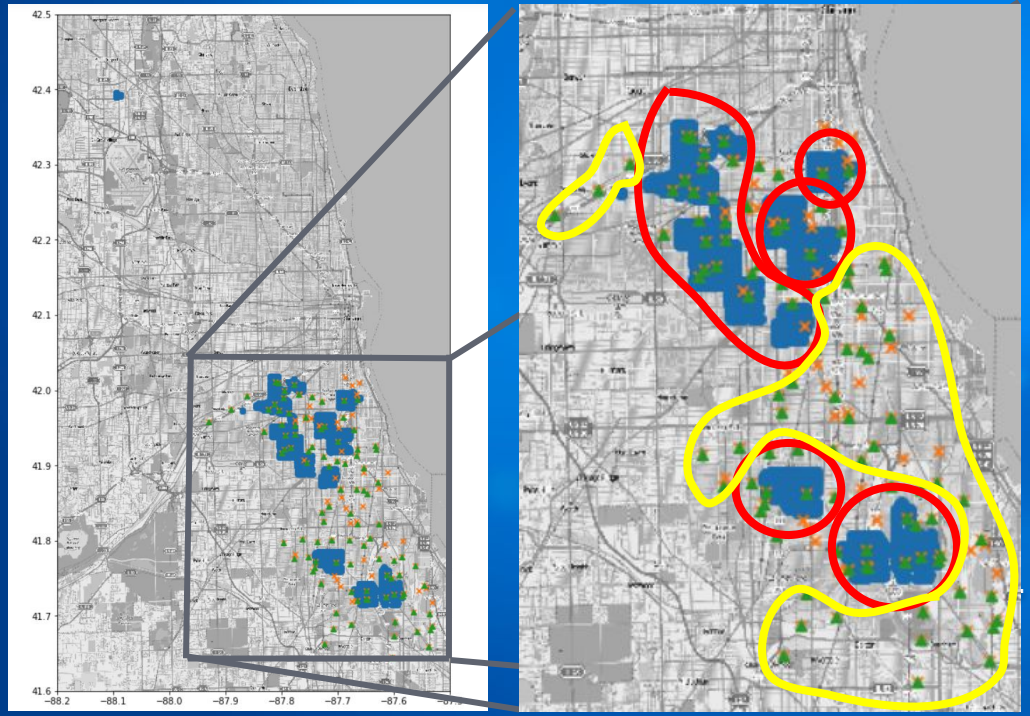


West Nile Virus has been affecting areas in Chicago and Efforts have been taken by the government to monitor the mosquito counts and also conduct spraying in intervals. However there has not been a significant drop in the cases. We shall investigate the factors that aid in predicting the virus and work on how these factors can be used to optimise the efforts of the Chicago Department of Health in eliminating this virus.

Data Pipeline

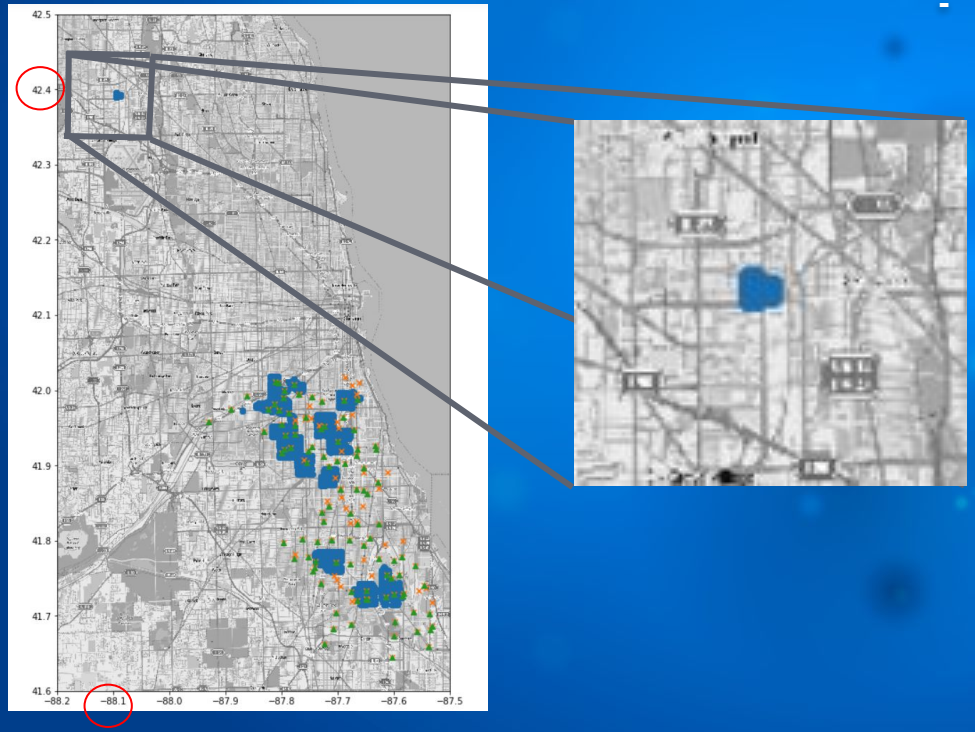


Train, Test and Spray Datasets



- Green triangles are areas where Wnv was present
- Orange X indicate traps deployed
- Blue blobs represent areas where pesticides were sprayed
- Spraying appears correlated over space (clusters)
- Not all Wnv-prone areas were covered (yellow clusters)

Train, Test and Spray Datasets



- "Outlier" cluster consisting of **95** spray instances conducted on 29 Aug 2011 at various time points

No traps nor evidence of Wnv present here

So why was spraying conducted there??

```
spray[(spray['Latitude']>42.3) & (spray['Longitude']<-88.0)]
```

	Date	Time	Latitude	Longitude
0	2011-08-29	6:56:58 PM	42.391623	-88.089163
1	2011-08-29	6:57:08 PM	42.391348	-88.089163
2	2011-08-29	6:57:18 PM	42.391022	-88.089157
3	2011-08-29	6:57:28 PM	42.390637	-88.089158
4	2011-08-29	6:57:38 PM	42.390410	-88.088858
...	...	...	...	...
90	2011-08-29	7:14:38 PM	42.392902	-88.093853
91	2011-08-29	7:14:48 PM	42.392587	-88.093867
92	2011-08-29	7:14:58 PM	42.392308	-88.093873
93	2011-08-29	7:15:18 PM	42.392183	-88.093767
94	2011-08-29	7:15:28 PM	42.392508	-88.093847

95 rows × 4 columns

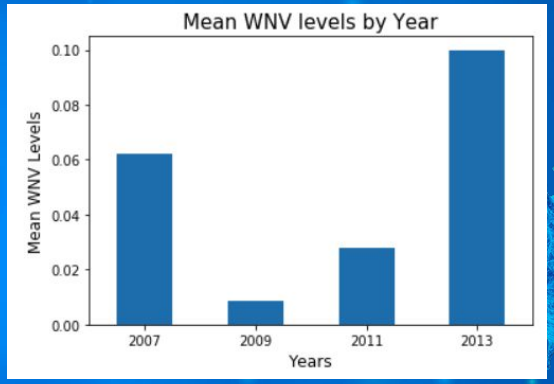
Train, Test and Spray Datasets

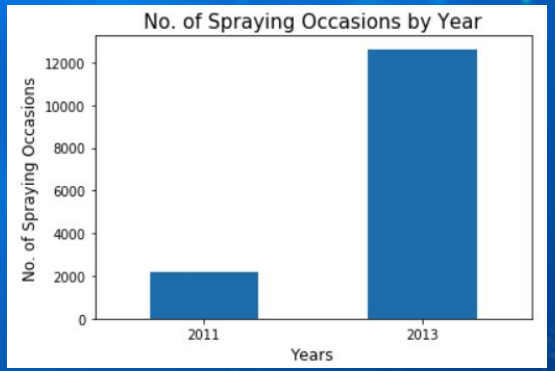
But it seems that..

- Spraying did not help curb WNV (no decline in Mean WNV from 2011 to 2013)
- Possibly a lot more spraying occasions in 2007 that led to significant drop in 2009 WNV but data not available at time of analysis to confirm

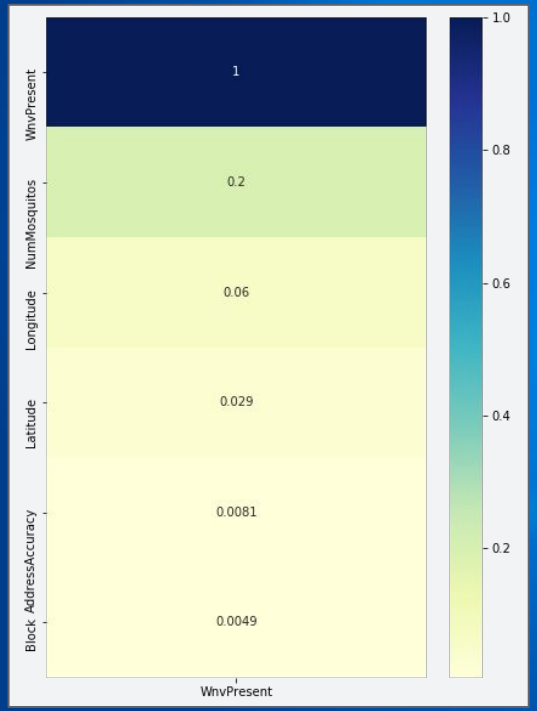
OR

- Less WNV in 2011 so less spraying; More WNV in 2013 thus more spraying
- **Insufficient spray data (only 2 / 8 years of spray data available)→ Will not be included in modelling**

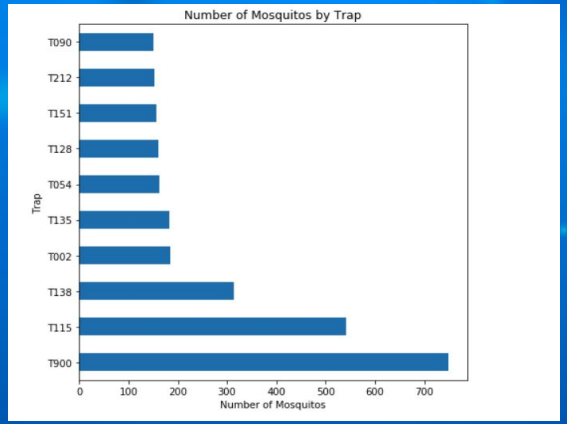


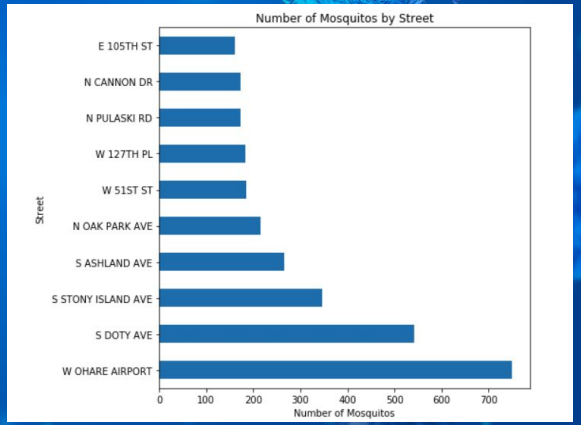


Train, Test and Spray Datasets

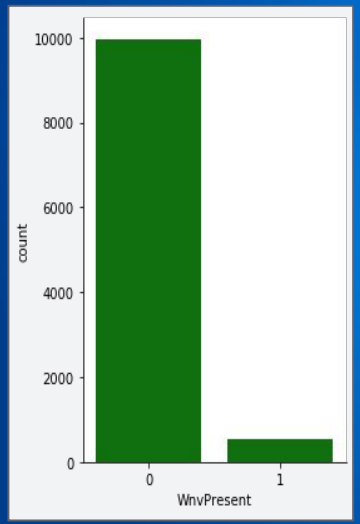


- Number of mosquitoes → Highest correlation with Target Variable
- Address related variables are unique to each Trap
- All address related variables except Longitude/Latitude shall be dropped



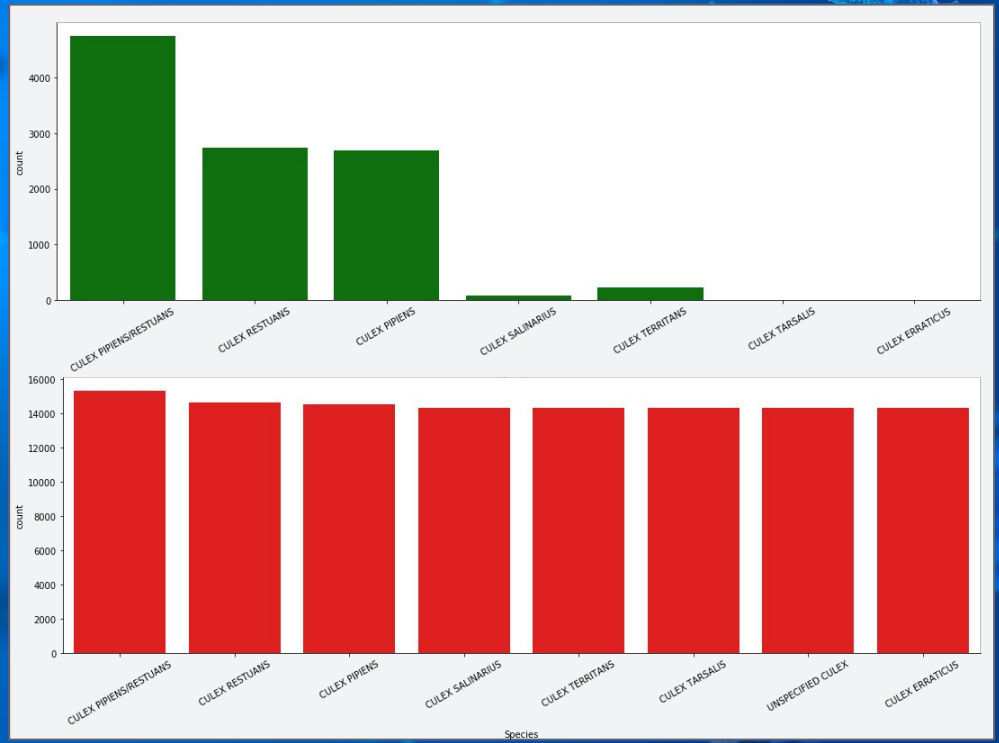


Train, Test and Spray Datasets

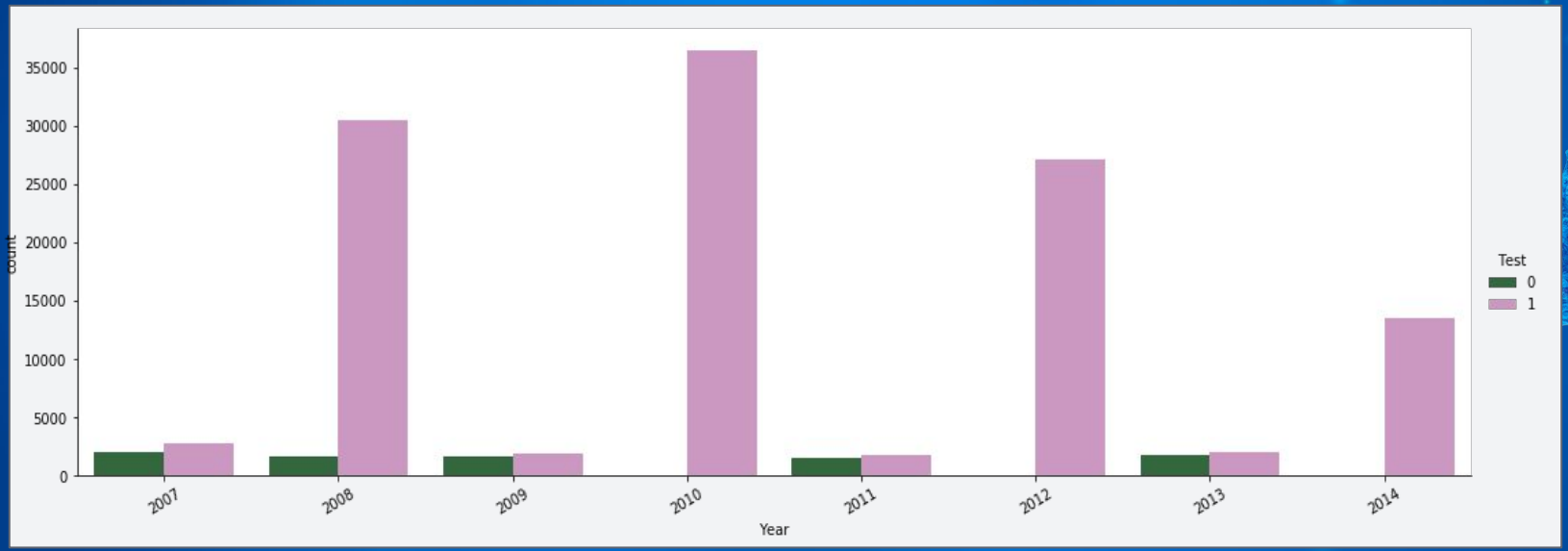


Imbalanced Data

- Overall positive case count was only about 5 %
- Distribution of mosquito species is not even within train data



Train, Test and Spray Datasets



- **Training data unavailable for 2010,2012 and 2014**
- **Month data shall be extracted and hot encoded**

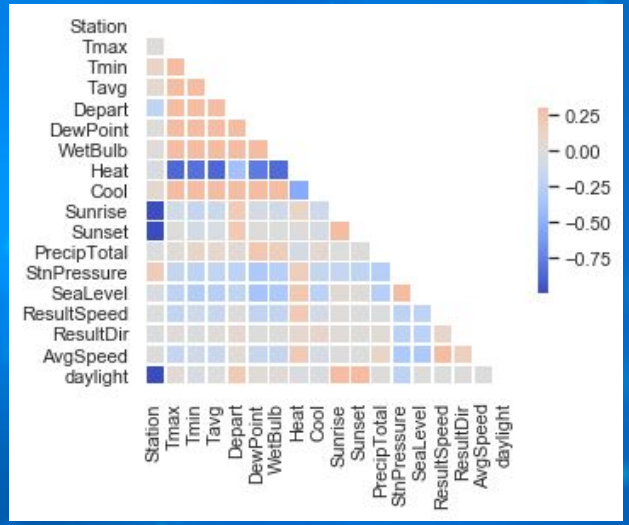
Weather EDA

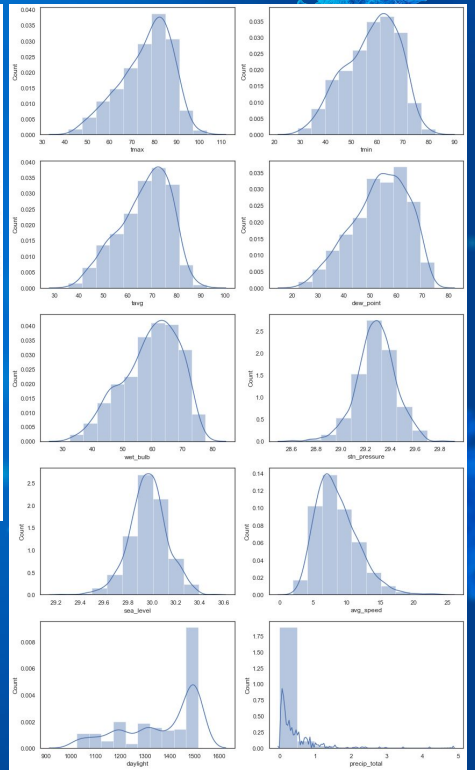
1. **Feature engineering:**
   a. **Created duration of daylight with Sunrise and Sunset**
   b. **Average values between Stations 1 and 2**

2. **Correlation present between features**

3. **Null values “M” or “-” are imputed with mean of respective features (<5% missing data)**

4. **Dropped features with high proportion of missing data**

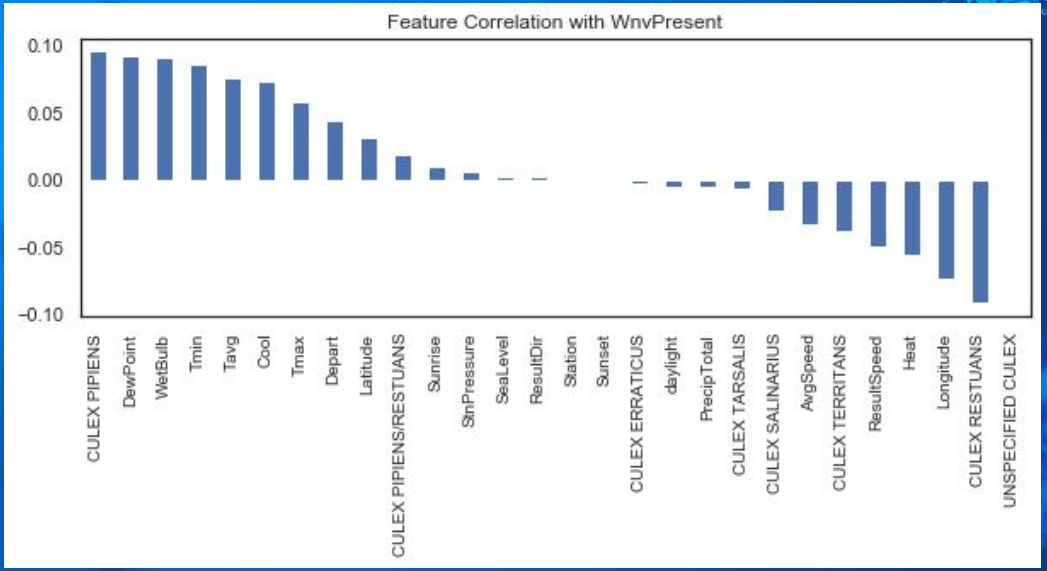




Combined train+weather EDA

Top features that are correlated to wnv presence:

1. **Mosquito species**
2. **Humidity (Dewpoint, WetBulb)**
3. **Temperature (Tmin, Tavg, Tmax)**



Modelling

Train-test split → Standard Scaler → Modelling

stratify = y
cv=5

train set: (8610, 166)
X_train: (6027, 166)
y_train: (6027,)
X_test: (2583, 166)
y_test is: (2583,)

```
y_train.value_counts(normalize=True)
```

```
0.0    0.946906
1.0    0.053094
Name: WnvPresent, dtype: float64
```

```
y_test.value_counts(normalize=True)
```

```
0.0    0.946961
1.0    0.053039
Name: WnvPresent, dtype: float64
```

GridSearch

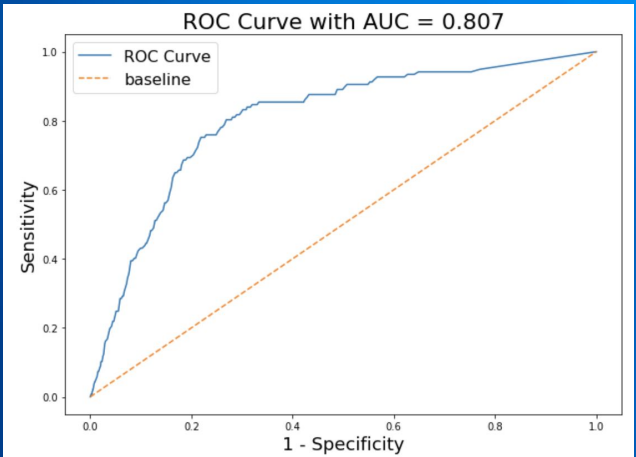
class_weights = balanced

Modelling

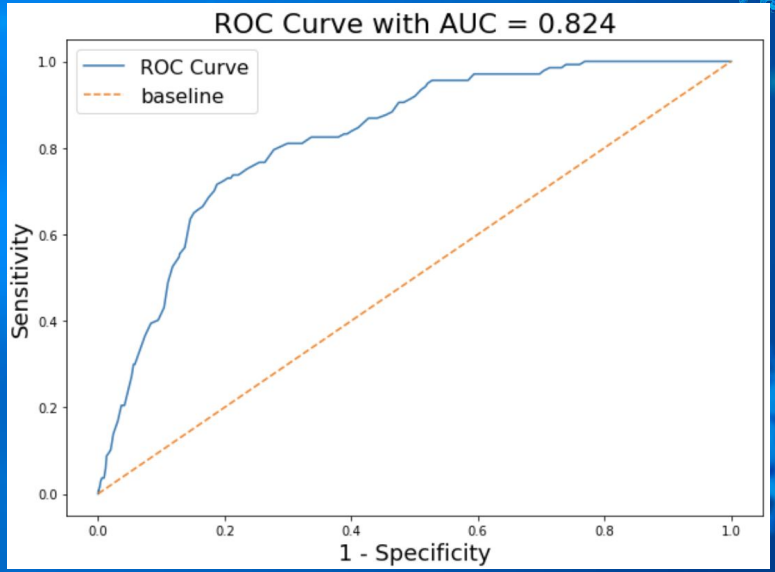
MODELS	train_accuracy	test_accuracy	specificity	precision	recall sensitivity	f1_score	roc_score
Vanilla Logistic Regression	0.9469	0.9458	0.9984	0.2	0.0073	0.0141	0.8132
Logistic Regression (GS)	0.746	0.7302	0.7261	**0.141**	0.8029	**0.2399**	0.8075
PCA + Logistic Regression (GS)	0.6789	0.6806	0.6791	0.11	0.708	0.1904	0.754
Random Forest (GS)	0.6811	0.688	0.6811	0.1246	**0.8102**	0.216	**0.8241**
XGBoost (GS)	0.6811	0.688	0.9984	0.2	0.0073	0.0141	0.8602

Modelling

Logistic Regression	Predicted no_Wnv	Predicted Wnv
Actual No_Wnv	1776	670
Actual Wnv	27	110

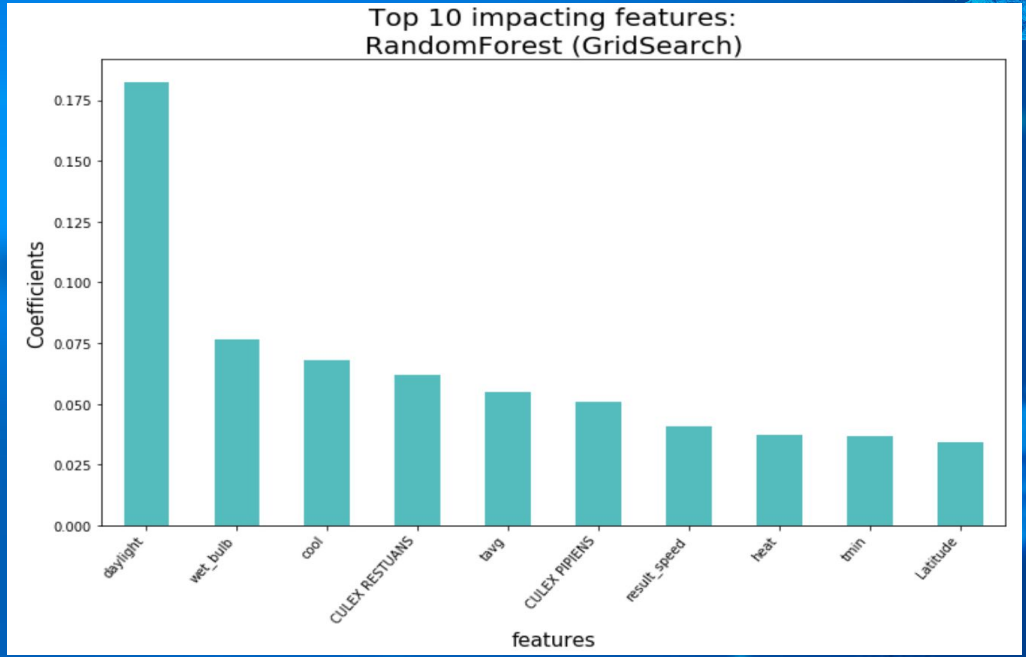


Random Forest	Predicted no_Wnv	Predicted Wnv
Actual No_Wnv	1666	780
Actual Wnv	26	111



Conclusion

- **High ROC Score**
- **RF handles imbalanced classes well**
- **Daylight feature has the highest coefficient → Longer day → Summer → Higher Temperature → Increased Wnv**



Limitations

- **The current spray technique does not seem to have a substantial impact thus it is not cost-effective.**
- **Birds form a essential part of the transmission cycle as they are the reservoir hosts and mosquitoes are merely the vectors.**

Recommendations

- Education:
  - Reiterate the importance of removing stagnant water to Farmers and wetland managers
  - Proper disposal of container like litter
  - Reduce outdoor activities during peak exposure time to mosquitos

- Surveillance:
  - Climatic Factors like Temperature play a crucial role → Visualise relationships to further make the assumption that high temperature and dry weather lead to high wnv cases
  - Increase surveillance from mosquitoes to Birds

- Mosquito Control
  - Elimination at larvae level is crucial as developing female mosquitoes will essentially multiply the population
  - Money can be directed towards developing effective repellents → block the mosquitoes' ability to smell humans → reduced transmission risk
  - Also the exact time of the spraying matter,The carriers are active early in the morning and in the evening.